«Невидимый веб» и некоторые вопросы доступности научной информации

Е. А. Негуляев,
главный библиотекарь
Научной библиотеки Уральского
государственного университета

Термин «невидимый веб» (invisible web) был впервые употреблен в 1994 году Джилл Иллсворт (Jill Ellsworth) для обозначения источников, по тем или иным причинам недоступных для обычных поисковых машин. В качестве синонимов используются «темный веб» (dark web) или «глубокий веб» (deep web). В противовес «невидимому» выделяется «поверхностный веб» (surface web), достижимый в результате простого перехода по ссылкам, соответственно — поддающийся индексации поисковыми системами.

Общий информационный объем невидимого веба в сотни раз превышает объем поверхностной части, при этом отношение числа «качественных» документов к их общему количеству тоже выше для невидимой зоны[1].

Основной причиной, по которой источники попадают в невидимую часть веба, является их интерактивный характер. Доступные через веб базы данных, динамически генерирующие информационные страницы после выполнения пользовательских запросов, не могут быть проиндексированными обычными поисковыми машинами, передвигающимися по ссылкам.

Важной составляющей невидимого веба является информация, размещенная не в традиционном для веб html-формате. Пионером среди мировых поисковых систем, распространивших индексирование и поиск за пределы html, был Google, который в феврале 2001 года стал индексировать файлы в pdf-формате. В ноябре этого же года Google добавил к списку поддерживаемых форматы Microsoft Word (doc), Microsoft Works (wks, wps, wdb), Microsoft Write (wri), Microsoft Excel (xls), Microsoft PowerPoint (ppt), Rich Text формат (rtf), PostScript (ps), Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku), Lotus WordPro (lwp), MacWrite (mw), Text (ans, txt). Эти события были восприняты как важные шаги по раскрытию невидимого веба, в том числе и научной его составляющей. Следует учитывать, что значительная часть документации, научных отчетов и статей представлена не в html-формате (чаще всего для этого используется pdf), поэтому ограничение области поиска по формату документов (например, pdf) способно обеспечить более «качественную» выдачу.

Чуть позднее, в мае 2002 года, документы в формате pdf стала индексировать поисковая система Fast (AllTheWeb), в настоящее время она поддерживает еще и поиск по Microsoft Word (doc) и Macromedia Flash (swf) файлам.

Для мировых поисковых машин поддержка нетрадиционных для веб форматов сейчас уже не является исключением, например, pdf файлы индексируют также Altavista и Scirus.

Из российских поисковых систем документы в нетрадиционных форматах первым стал индексировать Yandex. В феврале 2003 года[2] была добавлена поддержка pdf и rtf, а в июне введен поиск по документам в формате Microsoft Word (doc). На февраль 2003 года Yandex'ом

---

[1] Bergmann, Michael. The Deep Web: Surfacing Hidden Value [Электронный ресурс]. — 2001. — URL: http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/deepwebwhitepaper.pdf (проверено 20 сентября 2003 г.)

[2] URL: http://company.yandex.ru/news/2003/0220/ (проверено 20 сентября 2003 г.)

было проиндексировано около 50 тысяч документов в pdf формате[3], общее же количество их по оценкам представителей компании Rambler на этот момент достигало 706 тысяч[4] (с учетом pdf файлов, основанных на растровых изображениях и дубликатов).

В конце июля — первой половине августа 2003 года было проведено исследование с целью выяснения доступности и особенностей отражения русскоязычных pdf файлов через поисковые машины Google, Fast и Yandex. Все тесты проводились на протяжении нескольких дней в различное время, а полученные результаты усреднялись, чтобы избежать влияния случайных факторов. Тестовый период длился две недели и был закончен 11 августа, поэтому результаты экспериментов следует считать не изучением проблемы во всей ее динамике, а скорее снимком состояния на первую половину августа 2003 г.

Тестирование проводилось по следующей схеме:

1. определение общего количества русскоязычных pdf файлов, проиндексированных поисковыми машинами;
2. сравнение выдачи различных поисковых по группе тестовых однословных запросов с названиями научных отраслей и научными терминами[5];
3. Сравнение количества проиндексированных разными поисковыми машинами pdf файлов для некоторых избранных сайтов (т. е. глубина индексирования pdf на различных сайтах).

Для Google и Fast приведены усредненные данные, для Yandex, ввиду большого разброса результатов — раздельные результаты за несколько дней наблюдения:

	Google (ограничение выдачи только по русскому языку)	Google (база полностью)	Fast	Yandex (до 5 авг.)	Yandex (после 5 авг.)	Yandex 11 авг.
Общее количество русскоязычных pdf файлов	156 тыс.		12869	77 тыс.	96618	101299
Поисковый запрос:						
«математика»	1080	1463	354	479	616	647
«физика»	2112	2373	1208	888	1040	1100
«химия»	2560	2843	515	804	1035	1102
«литературоведение»	43	59	8	25	35	35
«минералогия»	108	144	26	33	40	42
«геология»	587	750	86	–	133	149
«география»	917	1029	130	348	497	506
Количество проиндексированных pdf файлов на отдельных сайтах:						
Уральское отделение РАН (http://www.uran.ru)	121	125	0	72	78	78

[3] URL: http://www.searchengines.ru/forum/showthread.php?s=ce9d44b8522b9e88641be00cfe81e859&threadid=3349 (проверено 20 сентября 2003 г.)

[4] URL: http://www.searchengines.ru/forum/showthread.php?s=ce9d44b8522b9e88641be00cfe81e859&threadid=3349 (проверено 20 сентября 2003 г.).

[5] При формулировке поискового запроса для Yandex'а ставилось условие на точное совпадение поискового термина, чтобы отсечь возможность поиска с учетом морфологии русского языка и поставить все поисковые системы в равные условия.

	Google (ограничение выдачи только по русскому языку)	Google (без ограничений)	Fast	Yandex (до 05.08.03)	Yandex (после 05.08.2003)	Yandex 11 авг.
Аудиториум: образовательный портал по гуманитарным наукам (http://www.auditorium.ru)	1765	1845	76	140	149	149
Международная соросовская программа в области точных наук (http://www.issep.rssi.ru)	850	1355	17	505	535	535
Институт автоматики и процессов управления Дальневосточного отделения РАН (http://www.iacp.dvo.ru)	6	21	0	0	0	0

Анализ результатов показывает, что база проиндексированных Google русскоязычных pdf документов на начало августа 2003 г. была приблизительно в 1,5 раза больше, чем у Yandex'а. Причем этот результат получается и при анализе общего количества проиндексированных pdf документов, так и при сравнении результатов реальных запросов. Глубина индексирования pdf файлов на отдельных русскоязычных научных сайтах у Google тоже оказывается больше. Такой результат был достаточно неожиданным, потому что при анализе по любым запросам без ограничения по формату файлов выясняется, что общая база русскоязычных документов Yandex'а больше, чем у Google. Вполне естественно, что национальная поисковая машина имеет преимущество над общемировой при индексировании русскоязычной информации, но непонятно, почему это преимущество не распространяется на pdf документы.

По сравнению с Yandex и Google поисковая машина Fast располагает намного меньшим количеством проиндексированных русскоязычных pdf документов: примерно в 10 раз меньше Google и почти в 8 раз меньше Yandex'а, но при этом обеспечивает достаточно высокую выдачу по тестовым запросам: всего в 1,5–2 раза меньше, чем Yandex. Объяснение этого факта требует специального исследования, возможно, Fast каким-то образом отдает предпочтение ресурсам с научной тематикой.

Yandex в течение тестового периода демонстрировал очень неравномерные результаты. В конце июля 2003 г. он располагал приблизительно 77 тыс. pdf документов, таким образом, пятимесячный прирост составил всего 27 тыс. файлов. После этого последовал «рывок» и пределах одного цикла обновления[6] было добавлено почти 20 тыс. файлов. Именно в первых числах августа у Yandex'а на короткий период был нарушен сложившийся цикл обновления[7], поэтому вполне возможно представить два этих события связанными между собой. Резкое повышение внимания этой поисковой машины к индексации pdf файлов вряд ли может быть объяснено без разработчиков Yandex'а.

Тем не менее, для самой популярной в России поисковой системы Yandex на первую половину августа 2003 г. достаточно большое количество pdf документов все еще оставалось в зоне невидимого веба. Как минимум 50 тыс. (вероятно, намного больше) находятся в зоне «поверхностного веба» и могут быть со временем проиндексированы. Хотя Yandex и уступает Google, отставание может быть ликвидировано уже в первой половине сентября 2003 г., если

[6] Обычно Yandex обновляет поисковый индекс дважды в неделю — по вторникам и пятницам.

[7] См. URL: http://www.searchengines.ru/forum/showthread.php?s=886bf6543aedb7d910d4a096b7a91757&threadid=4965 (проверено 20 сентября 2003 г.)

Yandex'у удастся удержатся на уровне около 5 тыс. новых pdf документов за один цикл индексации.

В последнее время сбором и предоставлением доступа к электронным ресурсам начали заниматься библиотеки. При этом основным интерфейсом доступа к ним являются электронный каталог библиотеки. Для связи электронного издания с библиографическим описанием используется стандартное для всех форматов семейства MARC поле 856, в котором записывается URL доступа. Сами же электронные документы чаще всего хранятся на веб- или ftp- сервере. Этот способ является стандартным и обладает большим универсализмом. Но так как технически любой электронный каталог представляет собой базу данных, работающую с пользовательскими запросами, находящаяся в нем информация не может быть проиндексирована обычными поисковыми машинами. В результате в поле невидимого веба попадают и все полнотекстовые документы, связанные с библиографическим описанием.

Следует отметить, что число доступных таким образом файлов оказывается достаточно большим. Например, Научная библиотека Уральского государственного университета (НБ УрГУ) предоставляет доступ к 184 авторефератам диссертаций и 10 полным текстам диссертаций, Научная библиотека Томского политехнического университета — к примерно 140 изданиям (авторефератам диссертаций и тезисам докладов на конференциях), а Научная библиотека Санкт-Петербургского государственного технического университета — примерно к 1600 авторефератам диссертаций.

Эти значительные ресурсы научного характера абсолютно неизвестны поисковым машинам. При этом единственная причина устраняется весьма легко — достаточно иметь страницу со ссылками на полные тексты, которая находилась бы в зоне поверхностного веба и могла бы быть проиндексирована поисковыми машинами.

Желание сделать свои электронные ресурсы доступными для поисковых машин было одной из причин построения еще одного интерфейса доступа к электронной коллекции авторефератов и диссертаций Научной библиотеки УрГУ, открытого 12 августа 2003 г. Доступ организован через сайт НБ УрГУ[8] и предоставляет собой автоматически формируемые списки, критерием отбора для которых является начальная буква фамилии автора или код специальности. Каждое библиографическое описание из списка связано с электронной версией в формате pdf. По сравнению с электронным каталогом этот способ является упрощенным, но при этом значительно более быстрым и удобным для некоторых задач. В дополнение ко всему он располагает все ссылки на полные тексты в зоне «поверхностного веба» и делает их потенциально доступными для поисковых систем.

Уже 26 августа значительная часть этой коллекции была включена в поисковый индекс Yandex'а[9].

Мы надеемся, что это упростит доступ к собранным нашей библиотекой научным ресурсам и позволит более эффективно решать задачу популяризации достижений ученых нашего университета.

В целом же следует отметить, что проблема попадания информации в невидимый веб зависит не только от особенностей поисковых машин, но и от особенностей публикации материалов. Любая организация, предоставляющая материалы в свободный веб-доступ, обязана принять дополнительные меры, если не желает, чтобы ее ресурсы были потеряны для многих пользователей веба.

---

[8] URL: http://lib.usu.ru/getetd.asp?id=0&table=etd (проверено 20 сентября 2003 г.)

[9] Подробнее о результатах индексации см.: Негуляев Е. А. Yandex и полнотекстовая коллекция авторефератов [Электронный ресурс]. — 2003. — URL: http://mlist.sgu.ru/pipermail/diglib/2003-August/000050.html (проверено 20 сентября 2003 г.)